BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

LƯƠNG DUY BẢO

NGHIÊN CỨU DECONVERTER ĐỂ DỊCH TỰ ĐỘNG TỪ UNL SANG TIẾNG VIỆT

Chuyên ngành: KHOA HỌC MÁY TÍNH

Mã số: 60.48.01.01

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng - Năm 2016

Công trình được hoàn thành tại
ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: PGS.TS. VÕ TRUNG HÙNG

Phản biện 1: TS. Đặng Hoài Phương

Phản biện 2: TS. Trần Thiên Thành

Luận văn đã được bảo vệ tại Hội đồng chấm Luận văn tốt nghiệp thạc sĩ Kỹ thuật họp tại Đại học Đà Nẵng vào ngày 25 tháng 7 năm 2016.

Có thể tìm hiểu luận văn tại:

Trung tâm Thông tin - Học liệu, Đại học Đà Nẵng

MỞ ĐẦU

1. Lý do chọn đề tài

Trong thời đại công nghệ thông tin với sự phát triển nhanh và đạt được nhiều thành tựu to lớn trong các lĩnh vực kinh tế, kỹ thuật, văn hoá, xã hội,… Cùng với sự phát triển này, nhân loại đã tạo ra lượng thông tin khổng lồ và phần lớn những thông tin đó chúng ta có thể tìm thấy thông qua hệ thống mạng Internet. Tuy nhiên, lượng thông tin này con người vẫn chưa được khai thác hết bởi rất nhiều lý do, lý do quan trọng nhất dẫn đến việc hạn chế khai thác thông tin trên, đó là rào cản ngôn ngữ vì nhiều con người chỉ biết mỗi ngôn ngữ mẹ đẻ hoặc biết một trong số những ngôn ngữ phổ biến như tiếng Anh, tiếng Pháp, tiếng Tây Ban Nha, tiếng Trung Quốc. Với sự phát triển của công nghệ thông tin giải pháp nhằm phá bỏ rào cản ngôn ngữ là phát triển các hệ thống dịch tự động.

Nhiều nghiên cứu về dịch tự động đã tạo những công cụ dịch hiệu quả và được sử dụng phổ biến như Google, AltaVisa,… Có những hệ thống thương mại hoá như Systran, Reverso, Babylon, … Những hệ thống này cho phép tạo ra "bản dịch nghĩa" – một bản dịch chưa được hoàn chỉnh nhưng giúp chúng ta có thể hiểu được ý nghĩa của văn bản gốc và cần phải chỉnh sửa nhiều để trở thành một văn bản hoàn chỉnh. Các hệ thống dịch tự động cho phép dịch rất nhanh và chi phí thấp hơn nhiều so với dịch bằng con người. Tuy nhiên, những hệ thống này đang phải đối mặt với rất nhiều vấn đề như sự đa nghĩa của từ, sự nhập nhằng về ngữ nghĩa, sự phụ thuộc về ngữ cảnh và nhiều khó khăn về sự khác biệt nhằm giải thích các khái niệm.

Hiện nay, nhu cầu về các hệ thống xử lý ngôn ngữ tự nhiên ngày càng tăng vì nó cần được ứng dụng trong nhiều lĩnh vực vì vậy vấn đề xử lý ngôn ngữ tự nhiên hiện nay rất cần các tài liệu song ngữ hoặc đa

ngữ, cho nên phát triển các hệ thống xử lý ngôn ngữ tự nhiên dựa trên các kho ngữ liệu đang được nghiên cứu và triển khai ngày càng nhiều.

Hiện nay, dịch tự động góp phần quan trọng trong quá trình phát triển kinh tế và xã hội của các quốc gia bởi nó xóa bỏ rào cản về ngôn ngữ. Trên thế giới có nhiều kho ngữ liệu song ngữ như Anh – Pháp, Anh – Hoa,… song việc nghiên cứu và phát triển các ứng dụng để dịch tự động từ UNL sang tiếng Việt còn hạn chế. Được sự gợi ý của PGS.TS Võ Trung Hùng tôi chọn đề tài ***"****Nghiên cứu DeConverter để dịch tự động từ UNL sang tiếng Việt****"*** làm đề tài tốt nghiệp luận văn cao học.

2. Mục đích và nhiệm vụ đề tài

Mục đích của đề tài tìm hiểu và trình bày tổng quan về dịch tự động, các phương pháp dịch trên cơ sở đó nghiên cứu hệ thống DeConverter để dịch tự động từ UNL sang tiếng Việt phục vụ việc xử lý ngôn ngữ tự nhiên liên quan đến tiếng Việt.

Nhiệm vụ cụ thể là nghiên cứu về ngôn ngữ UNL; nghiên cứu quá trình thực hiện chuyển đổi từ ngôn ngữ UNL sang ngôn ngữ tự nhiên – Tiếng Việt thông qua hai phần mềm DeCo và Eugene; sau khi thử nghiệm, đánh giá và đề xuất giải pháp ứng dụng các công cụ phát triển của UNL cho tiếng Việt. Kết quả nay sẽ tạo tiền đề cho việc nhanh chóng xây dựng thành công hệ thống dịch tự động đa nghĩa cho tiếng Việt trong tương lai.

Chúng tôi tập trung nghiên cứu về hệ thống UNL và tìm hiểu về dịch tự động, các phương pháp dịch tự động. Ngoài ra, nghiên cứu về kho ngữ liệu (linguistics corpus) và kho ngữ liệu đa ngữ (multilingual linguistics corpus).

Nghiên cứu sâu về hệ thống DeConverter trong UNL và các công cụ của nó như DeCo, Eugene.

Đề tài đề xuất giải pháp và thực nghiệm các công cụ dịch tự động sẵn có để dịch tự động các biểu thức UNL sang tiếng Việt từ các kho ngữ liệu sẵn có.

3. Đối tượng và phạm vi nghiên cứu

Trong khuân khổ của một luận văn thực nghiệm chúng tôi tập trung nghiên cứu hệ thống UNL: Ngôn ngữ UNL và mô hình hoạt động của hệ thống dịch UNL, kho ngữ liệu (Linguistics Corpus), các mô hình triển khai hệ thống và nghiên cứu kỹ các công cụ DeCoverter: Eugene, DeCo để thực nghiệm dịch các biểu thức UNL sang tiếng Việt. Ngoài ra, chúng tôi cũng giới thiệu tổng quát các hệ thống dịch hiện nay trên Internet như Systran, Google, Reverso,… trên cơ sở một số bài báo và luận văn tốt nghiệp khóa trước.

4. Phương pháp nghiên cứu

Phương pháp nghiên cứu, chúng tôi đã sử dụng hai phương pháp chính là nghiên cứu lý thuyết và nghiên cứu thực nghiệm. Với phương pháp nghiên cứu lý thuyết chúng tôi tiến hành thu thập các tài liệu về cơ sở lý thuyết: Hệ thống UNL – DeConverter, dịch tự động, các tài liệu mô tả một số công cụ DeConverter và các tài liệu liên quan đến vấn đề nghiên cứu.

Đối với phương pháp thực nghiệm chúng tôi nghiên cứu sử dụng các công cụ của DeConverter để dịch các kho ngữ liệu sẵn có sang ngôn ngữ tiếng Việt, sau đó thực nghiệm dịch và kiểm tra một số biểu thức UNL sang tiếng Việt.

5. Bố cục của luận văn

Bố cục của của luận văn được tổ chức thành 3 chương chính như sau:

Chương 1. Tổng quan về dịch tự động và UNL

Chương 2. UNL DeConverter và các công cụ

Chương 3. Thử nghiệm dịch biểu thức UNL sang tiếng Việt bằng các công cụ của DeConverter

Kết luận và hướng phát triển của đề tài.

CHƯƠNG 1
TỔNG QUAN VỀ DỊCH TỰ ĐỘNG VÀ UNL

Trong chương này, chúng tôi giới thiệu về lịch sử phát triển của dịch tự động, sau đó trình bày tổng quan về ngôn ngữ UNL và trình bày chi tiết quá trình mã hóa (EnConverter) và giải mã (DeCoverter) trong hệ thống UNL.

1.1. GIỚI THIỆU VỀ DỊCH TỰ ĐỘNG

1.1.1. Khái niệm

Dịch tự động hay còn gọi là dịch máy (MT: Machine Translation) là một nhánh của xử lý ngôn ngữ tự nhiên thuộc phân ngành trí tuệ nhân tạo, nó là sự kết hợp giữa ngôn ngữ, dịch thuật và khoa học máy tính. Dịch tự động thực hiện dịch một ngôn ngữ này sang một hoặc nhiều ngôn ngữ khác một cách tự động, có hoặc không có sự trợ giúp của con người [15].

Thuật ngữ dịch máy hay dịch tự động không chỉ bao gồm máy tính dựa vào các bộ từ điển và các phương pháp tiếp cận vào cơ sở dữ liệu để có thể đọc được văn bản hoặc tương tác với quá trình xử lý ngôn ngữ, chỉnh sửa văn bản để cho ra bản dịch cuối cùng, trường hợp bản dịch tự động hoàn toàn do máy tính không có sự can thiệp của con người có thể được hiểu như một "bản dịch nghĩa". Dịch tự động bằng máy tính nó không phải là lĩnh vực độc lập, mà nó là được kết hợp bởi các ngành khoa học khác để phát triển hoàn thiện.

1.1.2. Lịch sử phát triển

Những năm 50 thế kỉ XX, các nhóm nghiên cứu ở Hoa Kì và

Châu Âu đã tập trung vào lĩnh vực dịch tự động dưới sự bảo trợ của chính phủ và các nhà đầu tư. Đã có ít nhất ba hệ thống dịch tự động được nhiều tổ chức khoa học và quân sự thường xuyên sử dụng.

Những năm 60 thế kỉ XX, dịch tự động đã được nghiên cứu và thực hiện ở Hoa Kì, Canada và Châu Âu như Đại học Brigham Young (Hoa Kì) với hệ thống METEO, các nhóm GENA và SUSY (Châu Âu). Cuối những năm 70 thế kỉ XX, cùng với sự phát triển của kĩ thuật số, dịch tự động chuyển sang giai đoạn phát triển mới: dịch tự động được nghiên cứu và thực hiện với sự tham gia của con người vào các quá trình dịch của máy. Những năm 90, đã đánh dấu sự phát triển mạnh mẽ của công nghệ thông tin và kĩ thuật số trong nhiều lĩnh vực kinh tế và xã hội.

Hiện nay, nhiều sản phẩm dịch tự động đã và đang được sử dụng như các hệ thống dịch tự động với tốc độ dịch rất nhanh và độ chính xác khá cao. Dịch tự động (dịch bằng máy) được coi là mấu chốt trong các vấn đề kinh tế và xã hội của các quốc gia thời đại giao tiếp thông tin quốc tế.

1.1.3. Vấn đề dịch tự động cho tiếng Việt

Dịch tự động từ tiếng nước ngoài sang tiếng Việt được bắt đầu vào những năm 60. Tháng 6 năm 1970 hệ thống dịch tự động Logos I ra đời với từ điển tự động hóa hỗ trợ chỉ bao gồm hơn 1000 từ tiếng Việt. Năm 70 thế kỉ XX, hệ thống dịch tự động từ tiếng Anh ra tiếng Việt đã được tiến hành tại Tập đoàn viễn thông Xyzyx, California. Nhờ các hệ thống này, hiệu quả lao động của người dịch và chất lượng của cán bộ chuyên ngành được nâng cao. Việc dịch các đoạn văn dài, có nhiều từ chuyên ngành và phụ thuộc vào ngữ cảch thì độ chính xác chưa cao.

1.2. MỘT SỐ HỆ THỐNG DỊCH

1.2.1. Systran

Systran là một hệ thống dịch tự động rất nổi tiếng và có chất lượng dịch khá tốt được sử dụng trên môi trường Internet, có thể dịch được cho 52 cặp ngôn ngữ.

1.2.2. Reverso

Đây là hệ thống dịch tự động của Softissimo để dịch các văn bản hoặc các trang Web với định dạng HTML, có thể thực hiện dựa trên Internet, Intranet hoặc như là một ứng dụng trên máy đơn.

1.2.3. WorldLingo

WorldLingo là một công cụ dịch thuật miễn phí, nó có thể dịch một đoạn văn bản hay cả trang web. WorldLingo được kiểm nghiệm có độ chính xác dịch thuật lên đến 75%.

1.2.4. Google translate

Google Translate không chỉ mạnh ở khả năng dịch chuẩn, dịch sát ý, nó còn hỗ trợ nhiều ngôn ngữ, trong đó có tiếng Việt. Chức năng chính của Google Translate là dịch một câu, đoạn văn bản hoặc cả trang web sang ngôn ngữ yêu cầu. Google Translate sử dụng phần mềm dịch thuật riêng có chất lượng tốt, tuy nhiên nó chưa phải là công cụ dịch thuật hoàn hảo.

1.3. TỔNG QUAN VỀ NGÔN NGỮ UNL

UNL là từ viết tắt của "Universal Networking Language". UNL là ngôn ngữ máy tính cho phép máy tính có thể truy cập thông tin và tri thức mà không bị vấn đề về rào cản ngôn ngữ. Nó là ngôn ngữ có khả năng mô phỏng thế giới ngôn ngữ tự nhiên của con người trong giao tiếp. UNL biểu diễn thông tin hoặc tri thức dưới dạng mạng ngữ nghĩa với cấu trúc đa đồ thị. Sự biểu diễn của UNL là không nhập nhằng. Trong mạng đa ngữ nghĩa của UNL, nút biểu diễn khái niệm và

cạnh biểu diễn mối quan hệ giữa các khái niệm.

UNL bao gồm các thành phần của ngôn ngữ tự nhiên. UNL bao gồm Universal Word (UW) - từ vựng, Relation - Quan hệ, Attributes - Thuộc tính và UNL Knowledge Base - Cơ sở tri thức. Nó tạo ra các từ biểu diễn các khái niệm gọi là "Universal Word" gọi tắt là UW, UW chứa các từ vựng của UNL. Nó liên kết các từ vựng khác tạo thành câu. Những liên kết này gọi là "relation" - mối quan hệ, nó chỉ định vai trò của mỗi từ trong câu. Ngụ ý của người nói có thể được diễn tả thông qua "Attribute" - Thuộc tính.

"UNLKB" cung cấp định nghĩa về ngữ nghĩa của từ vựng. UNLKB định nghĩa quan hệ có thể có giữa các khái niệm, bao gồm các quan hệ phân cấp và kỹ thuật tham chiếu giữa các khái niệm, nghĩa của biểu thức UNL là không nhập nhằng. [10]

UNL có thể giảm chi phí phát triển tri thức và những khái niệm cần thiết cho quá trình xử lý tri thức bằng cách chia sẻ tri thức và khái niệm. Câu thông tin được biểu diễn thành đa đồ thị gồm có từ vựng (Universal Word - UW) gọi là nút và các quan hệ gọi là cạnh. Đa đồ thị này cũng được biểu diễn như là tập quan hệ nhị phân có hướng, mỗi quan hệ nhị phân chỉ mỗi quan hệ giữa hai từ vựng được biểu diễn trong câu. UNL biểu diễn thông tin phân thành chủ ngữ và vị ngữ. Chủ ngữ được biểu diễn sử dụng từ vựng và quan hệ. Vị ngữ được biểu diễn sử dụng thuộc tính được gắn kèm với từ vựng.

1.3.1. Biểu thức UNL

UNL biểu diễn thông tin và tri thức theo hình thái mạng ngữ nghĩa với đa đồ thị. Mạng ngữ nghĩa UNL là một tập các quan hệ nhị phân, mỗi quan hệ nhị phân bao gồm một quan hệ và hai từ vựng của mối quan hệ đó. Mối quan hệ nhị phân của UNL được biểu diễn theo dạng sau:

```
<relation>(<uw1>,<uw2>)
```

<relation> là một trong các quan hệ được miêu tả trong UNL Specifications.

<uw1>, <uw2> là hai từ vựng chứa trong mối quan hệ <relation>.

Đa đồ thị: Biểu thức UNL là mạng đa ngữ nghĩa. Một nút bao gồm mạng ngữ nghĩa của biểu thức UNL được gọi là "scope". Miêu tả dạng chung của quan hệ nhị phân của biểu thức UNL như sau:

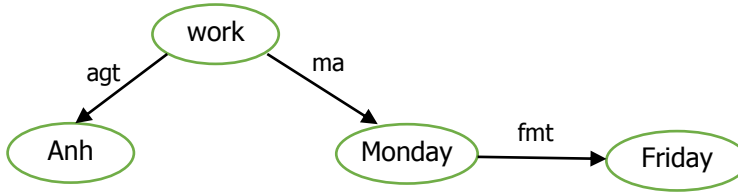
```
<relation>:<scope-id> (<node1>,<node2>)
```

Ví dụ 1: Ta có câu tiếng Anh: "Anh works from Monday to Friday"

Được biểu diễn trong UNL như sau:

```
agt(work(icl>do).@entry.@present,Anh(iof>person
))        ma(work(icl>do).@entry.@present,Monday)
fmt(Monday,Friday)
```

Được biểu diễn trong đa đồ thị như sau:



Hình 1.7. Biểu diễn đồ thị của ví dụ 1

1.3.2. Các quan hệ

Quan hệ của UNL được lựa chọn dựa theo các tiên đề sau:

Tiên đề 1: Điều kiện cần

Khi một từ vựng có quan hệ với nhiều hơn một từ vựng khác, mỗi quan hệ nên được gán nhãn và do đó có thể phân biệt các quan hệ

là điều kiện cần cho các khái niệm của mỗi biểu diễn UW.

Tiên đề 2: Điều kiện đủ

Khi có các quan hệ giữa các từ vựng thì nhãn quan hệ nên được gán làm sao để có thể hiện được vai trò của mỗi từ vựng.

Nhãn quan hệ được biểu diễn tối đa bởi chuỗi gồm ba ký tự. Quan hệ giữa các từ vựng là quan hệ nhị phân, các quan hệ có nhãn khác nhau dựa theo vai trò khác nhau mà chúng thể hiện.

1.3.3. Từ vựng UNL

a. Giới thiệu

Một từ vựng UNL không chỉ là một đơn vị cú pháp và ngữ nghĩa của UNL để diễn tả khái niệm, diễn tả một câu hoặc một khái niệm phức tạp. Từ vựng UNL được biểu diễn bằng một nút trong đa đồ thị của biểu thức UNL.

```
<UW> ::= <headword> [ <constraint list>]
<headword> ::= <character>…
<constraint list> ::= "(" <constraint> [ ","
<constraint>] … ")"
<constraint> ::= <relation label> { ">" | "<" } <UW>
[<constraint list>] | <relation label> { ">" | "<" }
<UW> [<constraint list>] ] …
<relation label> ::= "agt" | "and" | "aoj" | "obj" |
"icl" | …
<character> ::= "A" | … | "Z" | "a" | … | "z" | 0 | 1
| 2 | … | 9 | "_" | " " | "#" | "!" | "$" | "%" | "="
| "^" | "~" | "|" | "@" | "+" | "-" | "<" | ">" | "?"
```

Cấu trúc UNL: Một UW là một chuỗi các ký tự với các ràng buộc, gồm: Headword - Từ mục; Constraints or Retrictions - Ràng buộc…

b. Phân loại từ vựng UNL

Từ vựng là chuỗi ký tự, miêu tả các mức khái niệm khác nhau phụ thuộc vào các ràng buộc và có thể sử dụng để miêu tả các khái niệm đặc biệt, riêng biệt hơn, cụ thể hơn bằng cách đưa các thuộc tính và ID hoặc sự ràng buộc từ các miêu tả UNL khác. Có 4 loại từ vựng như sau:

Từ vựng cơ bản; Từ vựng ràng buộc; Từ vựng mở rộng; Từ vựng tạm thời.

Cách để trích dẫn từ vựng phức hợp

Được định nghĩa một lần, một từ vựng phức hợp có thể được trích dẫn hoặc tham chiếu đến bằng một cách đơn giản nhất là sử dụng ID của từ vựng phức hợp như là một từ vựng.

1.3.4. Thuộc tính từ vựng UNL

Thuộc tính của từ vựng được dùng để miêu tả chủ ngữ của câu. Nó cho thấy quan điểm và diễn đạt quan điểm của người nói.

1.4. HỆ THỐNG UNL

Một hệ thống UNL cơ bản bao gồm: UNL Language Servers, UNL Editor, UNL Viewer.

Khi một người bất kỳ truy cập vào Internet, người dùng có thể "mã hoá" (EnCoverter) văn bản được viết bằng ngôn ngữ của mình sang UNL. Tương tự, UNL bất kỳ có thể được "giải mã" (DeConverter) sang ngôn ngữ tự nhiên khác nhau. Quá trình xử lý "giải mã" và "mã hoá" được cung cấp bởi Languager Server được đặt trên mạng Internet. "Mã hoá" và "giải mã" có nhiệm vụ chuyển đổi một ngôn ngữ tự nhiên riêng biệt sang UNL và ngược lại.

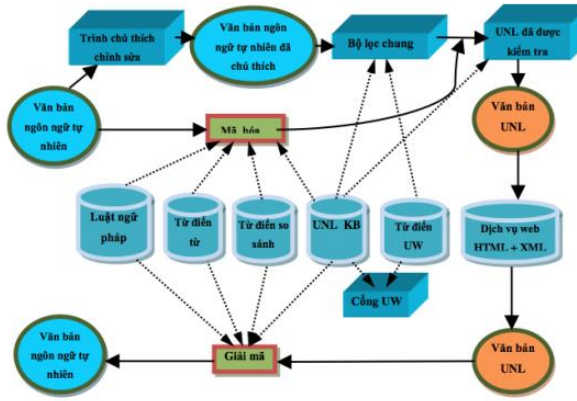
1.4.1. Quá trình EnConverter

Quá trình mã hoá (EnConverter) là phần mềm chuyển đổi một cách tự động văn bản từ ngôn ngữ tự nhiên sang UNL. Quá trình mã hoá làm việc theo cách sau: Chuỗi đầu vào sẽ được duyệt từ trái sang phải. Sau khi chuỗi đầu vào đã được duyệt xong, tất cả các hình vị được đánh dấu với các ký tự bắt đầu tương ứng trong từ điển và trở thành các hình vị thích hợp.

1.4.2. Quá trình DeConverter trong hệ thống dịch

Quá trình giải mã (DeConverter) được thể hiện trong hệ thống

dịch tự động được minh họa như sau:



Hình 1.16. Quá trình DeConverter trong hệ thống UNL

1.4.3. Từ điển – Dictionary

Từ điển là các từ lưu thông tin về ngôn ngữ, thông tin có liên quan đến loại của từ vựng, ngôn ngữ biểu diễn và nơi mà từ này có thể được sử dụng. Từ điển từ chứa các thành phần sau: UW cho định nghĩa khái niệm; Tiêu đề của từ để biểu diễn khái niệm; Thông tin về hoạt động cú pháp của từ.

CHƯƠNG 2
UNL DECONVERTER VÀ CÁC CÔNG CỤ

Trong chương này, chúng tôi giới thiệu tổng quan về DeConverter trong hệ thống UNL và các công cụ để DeConverter.

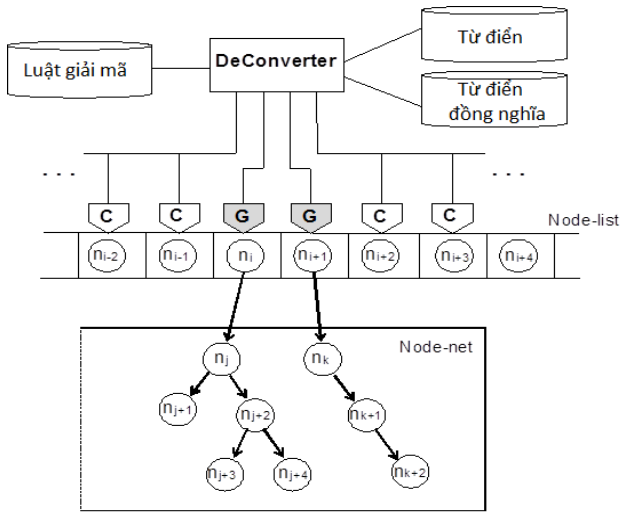
2.1. KHÁI NIỆM VỀ DECONVERTER

"DeConverter" là thành phần quan trọng trong hệ thống UNL, nó là phần mềm tự động giải mã UNL sang ngôn ngữ tự nhiên cho chất lượng dịch cao, kết quả chính xác. Nó là thành phần quan trọng là kiến trúc cơ bản của "DeConverter" để xử lý được tất cả các ngôn ngữ

với chất lượng và độ chính xác ở cùng một tiêu chuẩn. Công nghệ phát triển cho ngôn ngữ này có thể được áp dụng cho các ngôn ngữ khác miễn là các kiến trúc được chia sẻ. DeConverter dịch ra ngôn ngữ tự nhiên từ UNL, đóng vai trò nòng cốt trong hệ thống UNL.

Để có thể chuyển biểu thức UNL ra các ngôn ngữ tự nhiên khác chương trình cần sử dụng một số dữ liệu ngôn ngữ như từ điển, các luật về ngữ pháp và từ điển đồng nghĩa. Quá trình giải mã áp dụng các luật tổng quát chung cho từng nút trong Node - net và sinh ra danh sách từ trong ngôn ngữ đích. Quá trình giải mã là khả năng phát sinh tương tự như máy turing. Nó có khả năng phát sinh tất cả các loại câu, thích hợp đối với cả các loại ngôn ngữ. [14]

2.2. KIẾN TRÚC CỦA DECOVERTER



Hình 2.1. Kiến trúc của DeConverter

DeConverter hoạt động trên các nút của Node-list và chèn các nút từ Node-net vào Node-List thông qua các cửa sổ của nó. Có hai

loại cửa sổ, là Generation Windows và Condition Windows. Được viết tắt "Generation Windows" (GW), "Condition Windows" (CW). DeConverter tạo ra một câu bằng cách sử dụng từ điển - Word Dictionary, luật giải mã – DeConverter Rule và từ điển đồng nghĩa - Co-occurrence Dictionary. Generation Windows (GW) được sử dụng để kiểm tra hai nút liền nhau để áp dụng luật giải mã. Nếu có một nguyên tắc áp dụng, DeConverter sẽ sửa đổi các thuộc tính ngữ pháp của hai nút này và chèn một nút từ Node-net vào Node-list. Quá trình này sẽ được tiếp tục cho đến khi tất cả các nút của Node-net được chèn vào các Node-list, vì vậy mà các nút của Node-list sinh ra câu. Mỗi mục trong Word Dictionary được tạo thành từ ba yếu tố: các từ mục - Headword, từ vựng - Universal Word (UW) và các thuộc tính ngữ pháp. Một từ mục là một ký hiệu của một từ trong ngôn ngữ tự nhiên mà được sử dụng trong việc tạo ra một câu.

Một luật giải mã cho các nút gồm điều kiện đặt trên Generation Windows và Condition Windows. Giải mã bắt đầu từ nút entry và tất cả các nút của Node-net sẽ được chèn vào các Node-list khi thực hiện luật giải mã. Khi bắt đầu giải mã, nó chỉ bao gồm các nút entry của Node-net đưa vào. Mục đích của quá trình giải mã là để điền vào các Node-List với những từ thích hợp cho đầu vào của các Node-net. Khi một câu của biểu thức UNL được nhập trong khi làm Node-net, DeConverter sẽ truy vấn từ đến Word Dictionary cho mỗi node được sử dụng trong UW. Mỗi nút của Node-net sẽ bao gồm một danh sách các từ mục cần thiết, thể hiện như trong Hình 2.2.

Nút

Headword or part of a sentence notation	Universal Word	Grammatical Attributes

Hình 2.2. Cấu trúc của nút như một từ mục

Từ điển đồng nghĩa – Co-occurrence Dictionary cung cấp thông tin thực tế về từ ngữ của một ngôn ngữ bản địa.

2.3. CHỨC NĂNG CỦA DECONVERTER

DeConverter nhằm tạo câu mục tiêu của một ngôn ngữ có nguồn gốc từ biểu thức UNL bằng cách áp dụng luật giải mã. Nó kiểm tra định dạng của các luật và biểu thức UNL được nhập vào và các thông báo lỗi ở đầu ra. Khi sử dụng DeConverter bằng cách xây dựng các luật có thể dễ dàng phát triển và nâng cao luật.

2.3.1. Hoạt động

DeConverter tìm các nút trong Node-list thông qua các Generation Window và Condition Window, Khi nó tìm thấy một luật áp dụng, nó sẽ hành động hoặc hoạt động thông qua theo loại luật. Khi nó không tìm thấy bất kỳ luật áp dụng nào nó sẽ quay trở lại.

Công thức luật chèn bên phải:

```
:{<COND1>:<ACTION1>:<RELATION1>:<ROLE1>}
"<COND2>:<ACTION2>:<RELATION2>:<ROLE2>"
```

Mặt khác, để chèn một nút vào bên trái thực hiện theo luật sau:

```
:"<COND1>:<ACTION1>:<RELATION1>:<ROLE1>"{<COND2
>:<ACTION2>:<RELATION2>:<ROLE2>}
```

2.3.2. Word Dictionary Retrieval

Các biểu thức UNL được chuyển đổi thành các hình thức đầu vào của các Node-net, các mục từ điển cho tất cả UWs trong các biểu thức UNL được lấy từ điển của một ngôn ngữ mục tiêu. Quá trình Dictionary Entry Retrieval thay đổi tùy theo sự khác nhau của UW.

Mục từ điển: DeConverter cố gắng để lấy tất cả các mục từ điển cho mỗi UW từ Word Dictionary và sắp xếp theo thứ tự, theo các ưu tiên và độ dài của từ.

Lấy mục từ Dictionary Word có định dạng sau:

```
[HW]{} "UW" (ATTR1, ATTR2, …) <FLG, FRE, PRI>;
```

2.3.3. Tham chiếu một từ mục

2.3.4. Nguyên tắc áp dụng luật

Các quy tắc được áp dụng phụ thuộc vào Node-list mà Generation Windows được đặt ưu tiên các quy luật. Luật với cùng độ ưu tiên được áp dụng theo thứ tự chúng xuất hiện trong các tập tin văn bản của luật. Khi không có luật áp dụng được tìm thấy, nó sẽ quay lui. Khi hệ thống cố gắng áp dụng một luật để một nút mà không có từ mục được quyết định từ các từ tương ứng, hệ thống nhìn lên danh sách các từ tương ứng và tìm một lời đáp ứng các điều kiện của luật. Nếu một từ được tìm thấy, nó sẽ được lựa chọn và áp dụng luật thành công. Nếu không có từ được tìm thấy, việc áp dụng luật thất bại.

2.3.5. Quay lui

Quay lui xảy ra khi luật áp dụng không được tìm thấy hoặc luật áp dụng bị lỗi.

2.4. LUẬT GIẢI MÃ

Luật giải mã mô tả các điều kiện áp dụng luật: cách viết lại các thuộc tính của các nút đó đáp ứng những điều kiện, cũng như cách sáng tác một câu ngôn ngữ bản địa. DeConverter hoạt động trên các nút trong Node-list thông qua các cửa sổ của nó và các điều kiện và hành động của một luật giải mã được kết hợp với cửa sổ. Mỗi một phần của quy tắc thể hiện các điều kiện, hoặc các hành động trên, các nút lân cận trong Node-list theo thứ tự của Left Condiction Windows (LCW hay PRE), Genenation Window (LGW), Middle Condition Windows (MCW hoặc MID), Generation Window Right (RGW) và Right Condition Windows (RCW hoặc SUF).

2.4.1. Cú pháp của luật

Một luật giải mã có cú pháp như sau:

```
<TYPE>
["("<PRE>")" ["*"]]..."{" | """" [ <COND1> ] ":" [
<ACTION1> ] ":" [ <RELATION1> ] ":" [ <ROLE1>] "}"
| """"
["("<MID>")" ["*"]]...
"{" | """" [ <COND2> ] ":" [ <ACTION2> ] ":" [
<RELATION2> ] ":" [ <ROLE2> ] "}" | """"
["("<SUF>")" ["*"]]...
"P(" <PRIORITY> ");"
```

Biểu thức trong {} và "" được sử dụng để mô tả các luật cho các nút được chỉ định bởi Generation Windows. Trong khi {} được dùng để chỉ các nút hiện có trong các Node-list "," được sử dụng để mô tả các nút đó sẽ được chèn vào.

2.4.2. Mô tả các trường có điều kiện

```
  <PRE>|<MID>|<SUF>|<COND1>|<COND2>
::= ["^" ] <grammatical attribute> { ","
["^"] <grammatical attribute> }...
```

2.4.3. Mô tả các hành động của trường

Cú pháp chung của các thuộc tính ngữ pháp như sau:

```
<ACTION1>|<ACTION2> ::= { [ "+" ] | "-" }
<grammatical attribute>{ "," { [ "+" ] | "-" }
<grammatical attribute> }...
```

Các hành động "+" và "-" chỉ định việc bổ sung và xóa các thuộc tính ngữ pháp tương ứng.

2.4.4. Hướng quan hệ ngữ nghĩa

2.4.5. Hướng quan hệ đồng nghĩa

2.4.6. Mức độ ưu tiên của luật

`<PRIORITY>` mô tả thứ tự ưu tiên của luật

Các giá trị ưu tiên trong khoảng 0-255. Quy định không có một giá trị ưu tiên 0. Giá trị ưu tiên càng lớn thì mức ưu tiên cao

```
<PRIORITY>    ::= "0" | "1" | … | "255"
```

2.5. CÁC KIỂU LUẬT SINH

Phần này mô tả các loại luật sinh và chức năng của từng luật. Có 10 loại luật sinh, mỗi loại được đại diện bởi 1-2 ký hiệu như ':' '?' '? L', 'R', 'L',,, '? R', 'C', ' CR ',' DL 'và' DR 'đó sẽ được sử dụng trong <TYPE> trường quy định.

2.6. CÁC LUẬT HIỆU CHỈNH

Có 7 loại luật hiệu chỉnh, 'P >>', 'P <<', 'PR', 'PL', 'P:', 'PSR' và 'PSL', được sử dụng trong các <TYPE> lĩnh vực quy định. Nếu tùy chọn '^ pe' không có trong dòng lệnh thì các luật này được thực hiện, DeConverter sẽ áp dụng những quy tắc ngay sau khi các ứng dụng luật sinh kết thúc.

CHƯƠNG 3
THỬ NGHIỆM DỊCH BIỂU THỨC UNL SANG TIẾNG VIỆT BẰNG CÁC CÔNG CỤ CỦA DECOVERTER

Trong chương này, chúng tôi trình bày về quá trình thử nghiệm dịch các biểu thức UNL sang tiếng Việt bằng các công cụ DeCo và Eugene

3.1. CÔNG CỤ DECO

Công cụ DeCo hoạt động như sau: Đầu tiên biến đổi tập hợp các quan hệ nhị phân của biểu thức UNL đầu vào thành một cấu trúc đồ thị có hướng với các siêu nút gọi là node-net. Nút gốc của node-net được gọi là nút vào (entry node) là đầu của mỗi câu. Sau đó, áp dụng các luật giải mã đến các nút của node-net. Bắt đầu từ nút “entry node” để

tìm một từ thích hợp cho mỗi nút và tạo ra một chuỗi từ theo thứ tự ngữ pháp của ngôn ngữ tự nhiên. Quá trình giải mã kết thúc khi tất cả các từ cho tất cả các nút được tìm thấy và một chuỗi từ của câu đích được hoàn thành.

Thử nghiệm mẫu 1

- Biểu thức UNL đầu vào:

```
[S:1]
{org}
I ate rice
{/org}
{unl}
agt(eat(icl>consume>do,agt>living_thing,obj>con
crete_thing).@entry.@past,I(icl>person))
obj(eat(icl>consume>do,agt>living_thing,obj>con
crete_thing).@entry.@past,rice(icl>food))
{/unl}
[/S]
```

- Từ điển tiếng Việt – UNL cần thiết như sau:

```
[tôi]{}
"I(icl>person)"(PRON,HPRON,1SG,SUBJ)<v,0,0>;
[cơm]{} "rice(icl>food)"(N,SUBJ)<v,0, 0>;
[đã ăn]{}
"eat(icl>consume>do,agt>living_thing,obj>concrete_t
hing)"
(VT, S, 3SG.S,AGT.S,BA, BAE, OBJ.S,V,VDO,OBJ.S)
<v,0,0>;
```

- Xây dựng tập luật giải mã:

Luật 1: Điều khiển dịch chuyển cửa sổ

```
R {}{} P1;
```

Luật 2: Luật chèn đối tượng của quan hệ "agt" vào Node List

```
: "HPRON,SUBJ,1SG:subj:agt"
{V,^IRG,^pred:pred,1sg} P120;
```

Luật 3: Luật chèn đối tượng của quan hệ "obj" vào Node List

```
:{V,VDO,pred,^OBJ_inserted:OBJ_inserted}":obje:
obj" P100;
```

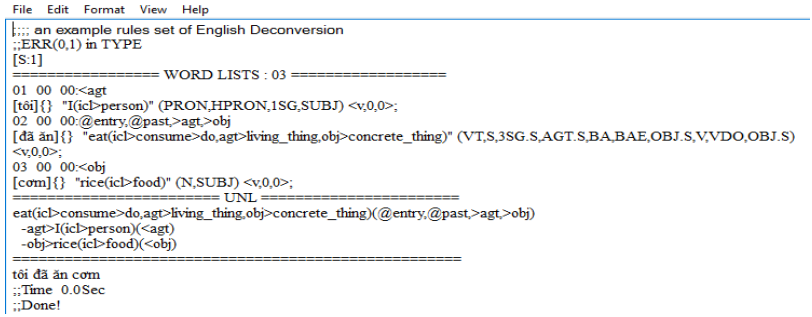
Luật 4: Chèn nút trống cho đại từ

```
: {PRON,^blk:blk} "[ ]:blk" P80;
```

Luật 5: Chèn nút trống cho động từ

```
: {V,^blk:blk} "[ ]:blk" P80;
```

Kết quả thực hiện chương trình:



Hình 3.4. Câu tiếng Việt được tạo ra tương ứng

Thử nghiệm mẫu 2.

- Biểu thức UNL đầu vào:

```
[S:1]
{org}
I wrote
{/org}
{unl}
agt(write(icl>communicate>do,agt>person,obj>info
rmation,cao>thing,ins>thing,rec>person).@entry.@
past,I(icl>person))
{/unl}
[/S]
```

- Từ điển tiếng Việt – UNL cần thiết như sau:

```
[tôi]{}
"I(icl>person)"(PRON,HPRON,1SG,SUBJ)<v,0,0>;
[viết]{}
"write(icl>communicate>do,agt>person,obj>informa
tion,ins>thing,rec> person)" (V,VT,SNGT)
<v,0,0>;
```

- Xây dựng tập luật giải mã

Luật 1: Điều khiển dịch chuyển cửa sổ

```
R {}{} P1;
```

Luật 2: Luật chèn đối tượng của quan hệ "agt" vào Node List

```
: "HPRON,SUBJ,1SG:subj:agt" {V@past, ED,^pred:
pred };
```

Luật 3:

```
: "HPRON,SUBJ,1SG:subj:agt" {V@past, ED,^pred:
pred };
```

```
[S:1]
============WORD LIST :02===========
01 00 00:<agt
[tôi]{} "I(icl>person)"(HPRON,PRON,SUB)
<Vie,0,0>;
02 00 00:@entry,@past,>agt
[đã viết]{}" write( icl>communicate > do,
agt > person, obj> information,ins> thing,
rec > person)" (V,VTSNGT) <Vie,0,0>;
============UNL===========
write(icl>communicate>do,agt>person,obj>info
rmation, ins>thing, rec
>person)(@entry,@past,>agt)
====================================
tôi đã viết
;;Time 0.0   Sec
;;Done!
```

Hình 3.5 Câu tiếng Việt được tạo ra tương ứng

3.2. CÔNG CỤ EUGENE

EUGENE (dEp-to-sUrface GENErator) là ứng dụng phát triển trên môi trường WEB để tạo ra các câu của ngôn ngữ tự nhiên từ biểu thức UNL được dựa trên trên từ điển UNL-NL và tập luật chuyển đổi. EUGENE có 7 thẻ Welcome, NL Input, Dictionaries, T-Rules, D-Rules, EUGENE và Compare. Thẻ "NL Input" cho phép người sử dụng cung cấp tài liệu dưới dạng các biểu thức UNL. Thẻ "Dictionaries" do người sử dụng cung cấp từ điển UNL-NL theo các đặc tả của UNL. Thẻ "T-Rules" cung cấp ngữ pháp chuyển đổi từ UNL sang ngôn ngữ tự nhiên. Thẻ "D-Rules" cung cấp định hướng ngữ pháp để chuyển đổi từ UNL sang ngôn ngữ tự nhiên, thẻ này dùng để điều khiển token và cải thiện kết quả chuyển đổi ngữ pháp. Thẻ "EUGENE" cho ra kết quả mã hóa là các câu hoàn chỉnh trong ngôn ngữ tự nhiên. Thẻ "Compare" là thẻ so sánh kết quả với các kết quả khác.

Thử nghiệm mẫu 3

- Biểu thức UNL đầu vào:

```
{unl}
mod(car(icl>motor_vehicle>thing).@entry.@def,beauti
ful(icl>adj,ant>ugly))
{/unl}
```

- Từ điển tiếng Việt – UNL cần thiết như sau:

```
[xe ô tô] {} "car(icl>motor-vehicle>thing)" (LEX=N,
POS=NOU, NUM=SNGT)<vie,0,0>;
[đẹp]
{}"beautiful(icl>adj,ant>ugly)"(LEX=J,POS=ADJ)<vie,0,0>;
```

+ Ta cần sử dụng các luật:

Luật 1: Loại bỏ quan hệ "mod", tạo quan hệ mới "NA"

```
mod(%x,N;%y,J):=NA(%x;%y,DIS=BEF);
```

Luật 2: Bỏ thuộc tính @def, tạo quan hệ mới "NS" với hai Uws "%x" và "chiếc" với thuộc tính là (LEX=D,+POS=ART)

```
(%x,N,@def):=(NS(%x,-
@def;%y,[chiếc],+LEX=D,+POS=ART));
```

Luật 3: Tạo khoảng trống giữa các UWs trong quan hệ "NA"

```
NA(%x;%y):=(%x,+>BLK)(%y,+>BLK);
```

Luật 4: Tạo khoảng trống giữa các Uws trong quan hệ "NS"

```
NS(%x;%y):=(%y,+>BLK)(%x,+>BLK);
```

Luật 5: Tạo khoảng trống giữa các từ trong câu

```
(%x,>BLK)(%y,^BLK,^PUT,^STAIL):=(%x,-
>BLK)("",+BLK)(%y);
```

+ Biểu thức tiếng Việt được tạo thành

```
### Please select the sentence to run from the list above ###

 [S:1]
 {org}
 the beautiful car
 {/org}
 {vie}
 chiếc xe ô tô đẹp
 {/vie}
 {unl}
 mod(car(icl>motor-vehicle>thing).@entry.@def,beautiful(icl>adj,ant>ugly))
 {/unl}
 [/S]
Dictionary Lookup Time 0 seconds, 4 milliseconds.
Disambiguation Time 0 seconds, 0 milliseconds.
Transformation Time 0 seconds, 12 milliseconds.
Total Time 0 seconds, 16 milliseconds.
```

Hình 3.6 Câu tiếng Việt được tạo ra tương ứng

Chúng tôi đã tiến hành dịch 50 câu tiếng Việt bằng cả hai công cụ đồng thời tiến hành so sánh với Google Translate, kết qủa được thể hiện như sau:

Công cụ sử dụng	Tỉ lệ thay đổi của câu	
UNL-> tiếng Việt với DeCo	38 (76%)	12 (24%)
UNL-> tiếng Việt với EUGENE	37 (74%)	13 (26%)

Qua các thí nghiệm trên, chúng tôi thấy rằng kết quả dịch của mỗi công cụ phụ thuộc nhiều vào chất lượng xây dựng tập luật cho từng câu dựa trên cấu trúc cú pháp tiếng Việt.

KẾT LUẬN

Việc nghiên cứu về UNL và các công cụ của nó để ứng dụng cho tiếng Việt mặc dù còn hạn chế nhưng cũng đã có những thành công nhất định. Kết quả lớn nhất mà chúng tôi đã đạt được qua đề tài là đã nghiên cứu, trình bày một cách có hệ thống về UNL, nghiên cứu một số công cụ hỗ trợ của UNL và cũng đã giới thiệu các nghiên cứu thực hiện để ứng dụng cho tiếng Việt. Trong quá trình thử nghiệm chúng tôi đã tiến hành thử nghiệm dịch 50 biểu thức UNL sang tiếng Việt và đánh giá chất lượng giải mã các công cụ của UNL như: DeCo, EUGENE.

Thông qua các thí nghiệm, chúng tôi thấy rằng các công cụ chuyển đổi từ các biểu UNL sang tiếng Việt có chất lượng là chấp nhận được. Tuy nhiên, để sử dụng những công cụ này, người sử dụng phải tuân theo các định dạng dữ liệu quy định. Nhưng vẫn còn một số ngôn ngữ mà không tuân theo các quy tắc, do đó cần có những công cụ điều chỉnh. Để sử dụng được các công cụ này, ta cần có bộ từ điển tiếng Việt - UNL và cần có kiến thức chuyên môn về ngôn ngữ UNL, để từ đó có thể xây dựng được hệ thống tập luật cho từng câu, đây là một trong những khó khăn khi sử dụng những công cụ này.

Trong luận văn này, chúng tôi giới thiệu một cách tổng thể về UNL và dịch tự động, đồng thời nghiên cứu UNL trong dịch tự động và quá trình hoạt động của các công cụ ứng dụng UNL để dịch các

biểu thức UNL sang tiếng Việt. Chúng tôi cũng đã chứng minh làm thế nào để chuyển đổi thành một biểu thức UNL sang tiếng Việt. Mặc dù chúng tôi đã làm việc với các ví dụ mẫu, không phải là tất cả các câu tiếng Việt mà đó chỉ là sự khởi đầu của một dự án lớn. Đây cũng là một trong những bước đi để phát triển một hệ thống dịch đa ngôn ngữ cho tiếng Việt nói chung để hòa nhập với cộng đồng UNL trên thế giới, cụ thể hơn đó là tiền đề để có thể xây dựng một Server dịch tự động cho ngôn ngữ của các dân tộc Việt Nam, chẳng hạn như ngôn ngữ Chăm, Khmer và Cơ-tu tạo thành một hệ thống dịch tự động đa ngôn ngữ hoàn chỉnh được sử dụng ở Việt Nam.

Để đạt được mục tiêu đó việc nghiên cứu quá trình DeCoverter trong hệ thống UNL cũng là nội dung quan trọng, ngoài ra việc nghiên cứu phát triển hệ thống từ điển Việt-UNL, hệ thống mã hóa từ tiếng Việt sang UNL và hệ thống tập luật cũng cần được nghiên cứu và phát triển đồng bộ.